Niklas Lauffer

✉ nlauffer@berkeley.edu 🏠 niklaslauffer.github.io 🎓 niklaslauffer niklaslauffer

Education

University of California, Berkeley **2021 — May 2026 (expected)**

PhD in Computer Science (Artificial Intelligence)

Advised by Stuart Russell and Sanjit Seshia – NSF Fellowship – CAIF Fellowship

University of Texas at Austin **2017 — 2021**

BS in Computer Science Honors and Mathematics

Dean's Honored Graduate (awarded to the top 1% of graduates) – Turing Scholars honors – 3.96/4.0 GPA

Selected Publications

ICLR 2026	**Imitation Learning for Multi-Turn LM Agents via On-policy Expert Corrections** – **In submission** N. Lauffer, X. Deng, S. Kundurthy, B. Kenstler, J. Da
NeurIPS 2025	**Robust and Diverse Multi-Agent Learning via Rational Policy Gradient** N. Lauffer, A. Shah, M. Carroll, S. Seshia, S. Russell, M. Dennis
Arxiv 2025	**Multi-Agent Risks from Advanced AI** Hammond et al.
AAMAS 2025	**Learning Task Decompositions for Multi-agent Teams** A. Shah*, N. Lauffer*, T. Chen*, N. Pitta*, S. Seshia
NeurIPS 2024	**Compositional Automata Embeddings for Goal-Conditioned Reinforcement Learning** B. Yalcinkaya*, N.Lauffer*, M. Vazquez-Chanlatte, S. Seshia
arxiv 2024	**Welfare Diplomacy: Benchmarking Language Model Cooperation** G. Mukobi, H. Erlebach, N. Lauffer, L. Hammond, A. Chan, J. Clifton
NIPS GCRL 2023	**Automata Conditioned Reinforcement Learning with Experience Replay** – **Spotlight** N. Lauffer*, B. Yalcinkaya*, M. Vazquez-Chanlatte, S. Seshia
ICML 2023	**Who Needs to Know? Minimal Knowledge for Optimal Coordination** N. Lauffer, A. Shah, M. Carroll, M. Dennis, S. Russell
TAC 2023	**No-regret Learning in Dynamic Stackelberg Games.** N. Lauffer, M. Ghasemi, A. Hashemi, Y. Savas, and U. Topcu.
JAIR 2023	**On Expected Value Strong Controllability.** N. Lauffer, W. Lassiter, and J. Frank.
FMCAD 2022	**Deterministic Finite Automata Decompositions from Examples and Demonstrations** N. Lauffer, B. Yalcinkaya, M. Vazquez-Chanlatte, A Shah, S. Seshia
Automatica 2021	**Training Classifiers for Feedback Control with Safety in Mind.** H. Poonawala, N. Lauffer, and U. Topcu
COCOA 2020	**Reachability Games for Optimal Multi-Agent Scheduling of Tasks with Variable Durations** D. Raju, N. Lauffer, U. Topcu.
ICAPS XAIP 2019	**Human-Understandable Explanations of Infeasibility for Resource-Constrained Scheduling Problems** N. Lauffer, and U. Topcu
CDC 2018	**Expedited Learning in MDPs with Side Information** M. Ornik, J. Fu, N. Lauffer, K. W. Perera, M. Alshiekh, M. Ono, and U. Topcu

Work Experience

Center for Human-Compatible AI, Learn & Verify | UC Berkeley **2021 — Present**

PhD Candidate in Artificial Intelligence

My PhD is centered around AI safety, human-AI collaboration, multiagent reinforcement learning, and LLM agents.

- Developed a framework for generalizing adversarial learning algorithms to the cooperative and general-sum setting.
- Developed novel pretraining method for learning representations of multi-step-plans, frameworks for evaluating LLM capabilities, a library for decomposing formal specifications, more efficient LLM architectures in long-context settings, and wrote the "Coordination" section of the *Multi-Agent Risks from Advanced AI* report.
- Published eight first-author papers at top venues (NeurIPS, ICML, AAMAS, JAIR, FMCAD, TAC) with more co-authored.
- On the program committed for CHAI 2024 and CHAI 2025 which was attended by over 200 researchers. Led and organized all-hands meetings, discussions, and talks for CHAI from 2023-2025. This amounts to 140+ talks from external and internal researchers.

Scale AI | Reasoning and Agents Team — Summer 2025

Research Scientist Intern in LLM Agents

Interned on the Reasoning and Agents team, researching how to improve LM agent training for long-horizon, multi-turn tasks.

- Developed a novel training scheme for multi-turn LM agents that combats the issue of covariate shift.
- Improved the state-of-the-art 7B LM agent performance on SWE-bench from 12% → 20% and 32B from 36% → 40%.
- A full conference paper is under review at ICLR.

Autonomous Systems Group | UT Austin — 2017 — 2021

Student Researcher in Autonomous Systems

Advised by Ufuk Topcu in the Institute for Computational Engineering and Science. As part of the Autonomous Systems Group, I developed formal and empirical approaches to decision making (MDPs, planning, RL) and control for autonomous systems.

- Published 5+ papers at top venues (CDC, ICAPS, COCOA, ACC, Automatica, Scientific Reports).
- Developed the first no-regret algorithm for learning in dynamic Stackelberg games resulting in a first-author publication at TAC.
- Built quadcopter flight software in C/C++, reinforcement learning visualization tools and a neural controlled UAV in Python.

NASA Ames Research Center

Research Intern in Planning and Scheduling — Summer 2019 and 2020

Interned at NASA Ames Research Center in the Automated Planning and Scheduling group under Dr. Jeremy Frank.

- Project resulted in a first-author journal publication at JAIR.
- Formulated the theory behind rescheduling policies for Expected Value Probabilisitic Simple Temporal Networks (EvPSTNs).
- Implemented dynamic rescheduling simulations for EvPSTNs to evaluate the effectiveness of different rescheduling policies.

Academic Service

Organizer	CHAI Workshop 2024-2025, CHAI Internship 2023-2025, CHAI All-hands 2023-2025, PSBAI NSF Workshop 2022.
Reviewer	NeurIPS 2025, ICML 2025, RLC 2025, TAC 2024, ACC 2024, AAAI 2024, CAIF Grant Making 2024-2025
Advising	Darius Muglich, Rupali Bhati, Mariana Meireles, Sandy Tanwisuth, Martín Soto, Thomas Chen, Nikhil Pitta
Teaching	CS188: Artificial Intelligence (2022), CS370: Homotopy Type Theory (2020)

Honors

2024	**Cooperative AI Foundation Fellowship:** Fellowship to support research in Cooperative AI
2023	**NSF Graduate Research Fellowship:** Awarded to high-potential PhD students early in their career
2022	**Hertz Fellowship Finalist:** One of 42 finalists selected from over 750 applicants
2021	**University of Texas Dean's Honored Graduate** - highest honor awarded to 1% of graduating students
2021	**Turing Scholars (Computer Science Honors)** - less than 7% of students are admitted
2021	**Dean's Scholars (Math Honors)** - less than 1.5% of students are admitted
2021	**Turing Scholars' Best Undergraduate Thesis Award Finalist**

Invited Talks

Aug 2023	**University of Maryland MARL Group**: Who Needs to Know? Minimal Knowledge for Optimal Coordination
Aug 2023	**Berkeley Multi-agent Learning Seminar**: Who Needs to Know? Minimal Knowledge for Optimal Coordination
Aug 2023	**MIT Algorithmic Alignment Group**: Who Needs to Know? Minimal Knowledge for Optimal Coordination
Jul 2023	**ICML 2023**: Who Needs to Know? Minimal Knowledge for Optimal Coordination
Jun 2023	**CHAI Workshop 2023**: Who Needs to Know? Minimal Knowledge for Optimal Coordination
Dec 2022	**Nissan Alliance Innovation Lab**: Learning DFA Decompositions from Examples and Demonstrations.
Oct 2022	**FMCAD 2022**: Learning DFA Decompositions from Examples and Demonstrations.

Technical Skills

Languages (Advanced)	Python, C, C++, Java, Bash, LaTeX
Languages (Basic)	MATLAB, R, Haskell, z/OS Assembly, HTML, CSS
Libraries (Python)	Jax, Pytorch, NumPy, SciPy, Scikit-Learn, Gym, Matplotlib, Seaborn, DGL, ROS, Gurobi, Z3
Foreign Languages	German (Fluent Reading, Fluent Speaking, Fluent Listening, Intermediate Writing)